Toward Environment–to–Environment (E2E) Affective Sensitive Communication Systems

Marco Paleari
Multimedia Department
EURECOM
Sophia Antipolis, France
paleari @ eurecom.fr

Vivek Singh
Department of Information and Computer Science
University of California, Irvine
sigh @ uci.edu

Benoit Huet
Multimedia Department
EURECOM
Sophia Antipolis, France
huet @ eurecom.fr

Ramesh Jain
Department of Information and Computer Science
University of California, Irvine
jain @ ics.uci.edu

ABSTRACT

In the last decades several efforts have been focused on helping people keeping in touch and communicating from distant environments. Nevertheless, existing devices limit natural human interactions in a number of different ways. Researchers have been working on this direction and new systems allow to communicate in more natural ways.
A communication modality which has still not be fully considered is emotions. Emotions are proved to be fundamental for various human cognitive ability and in particular for communications. Indeed, it has been proved that the surrounding emotional information can, sometimes, be more important that the message itself.
In this work we present a framework for environment–to–environment communications (E2E) which also take into account emotions to dynamically modify the behavior of the system.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human Factor—*Human Information Process*; D.2.11 [**Software Architectures**]: Domain Specific Architectures; H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms

Human Factors, Algorithms, Design

Keywords

Emotion recognition; facial expressions; affective computing; telepresence; event; environment; human centered computing.

MTDL'09, October 23, 2009, Beijing, China.

1. INTRODUCTION

The ability to recognize emotions is intrinsic in human beings and is known to be very important for natural interactions, decision making, memory, and other cognitive functions [1, 2]. As an example, during face to face meeting, it has been suggested that as much as 93% of what we communicate when talking directly with others can be transferred through paralanguage (e.g. voice tone and volume, body language, facial expressions, etc.) [3].

Despite this fact, human communications are moving from person–to–person communication to device–to–device communications. Still, in telecommunications the limitations of the communication link and of the device do not generally allow users to transfer all the information that it is needed spontaneously. Indeed, existing devices have restricted the affordances available to the users in terms of physical movement [4, 5, 6], interaction, peripheral vision [7], spatio–semantic integrity, and, therefore, information flow [8].

Since 1980s researchers have experimented with connecting remote environments in the form of media spaces [9, 10, 11] in order to improve the communication quality. Media spaces in general use a combination of audio, video, and networking to create a 'virtual window' across a distance and into another room. However, the combination of technologies typically used in media spaces restricts naturalistic behavior [10]. Recently, there have been even more efforts at enhancing the feeling of co–presence across physical space, either by using specially fabricated meeting rooms which look like mirror images of each other (e.g. HP:HALO[12]), or exploring the other extreme of moving all the communication to the virtual world (e.g. SecondLife[13]).

Still, a part that can be semantically meaningful and fundamental for the information transfer is missing. This part is represented by emotions. In the last decade, researchers have approached the topic of automatic, computer based, emotion recognition [14, 15] in the attempt to make human–computer interactions and human telecommunications more similar to human–human interactions. Few works (MAUI [16]) have planned to include affect as a source of information for natural communications but none has ever developed a generic system for computer mediated interactions taking emotions into account.

To understand why emotions can be fundamental in environment to environment communications we could just think at the scenario of a professor teaching to a class of remote students. In this scenario, the ability to detect the state of boredom is fundamental to trigger a reaction from the part of the professor. Connection bandwidth and screen definition does not allow this information to pass via the video streams. From these simple observations we can conclude that some other paradigms have to be found in order to trigger the professor attention to students who might be getting off track.

All of these options present limits removing us from the grounded reality of natural environments in which we would ideally like to interact (HALO, SecondLife) or not putting the focus on allowing people to interact in general scenarios and in natural ways (MAUI).

In this work we propose EEE2E (Emotion Enhanced Environment 2 Environment) as the new form of communication which allows users to connect their natural physical environments for communications.

In EEE2E, multiple heterogeneous sensors, devices and technologies are used. However, their abundance and the underlying design architecture push them into a supporting role in the background to maintain the focus on natural *human–human–interaction*[1]. Emotions are recognized and the information about the affect of the user interacting is transferred to the remote machines to be used accordingly to the application scenario.

The paper is organized as follows: section 2 describes the general E2E architecture, section 2.1 overviews the algorithms involved in the emotion recognition part of this work, section 3 describes the phases involved in integrating the emotions in E2E system and the actual implementation; finally, section 7 reports some concluding remarks and leads for future work.

2. SYSTEM ARCHITECTURE

At first look, EEE2E system might look comparable to other video–conferencing or tele–immersive environments. However, videoconferencing/telepresence systems like HP's Halo [12], Microsoft's Roundtable, and Cisco's Telepresence support bidirectional interactivity but are totally oblivious to the situations (i.e. semantics of the multimodal content) they connect. Hence they result in systems which are rigid in terms of required set–up (e.g. specially crafted meeting rooms), applications supported, and bandwidth required. On the other hand, tele–immersive and Multi–perspective imaging works [17, 18] often understand user objectives to support enhanced user interaction, but they do so only unidirectionally. EEE2E communication systems support enhanced user affordances bi–directionally based on a semantic understanding of the environments connected. This has been illustrated in figure 1.

We extend the architecture described in [19] for emotion–enhanced E2E systems which abstract environment semantics based on emotions being portrayed. We continue to adopt an event–centric approach as the changes in emotions

[1]Thus, the users need not to worry about staying within proximity, field of view, audible distance and so on of a sensor or an output device (e.g. screen, speaker etc.) but rather just interact in their natural settings and let the system find the most appropriate input and output devices to support communication.

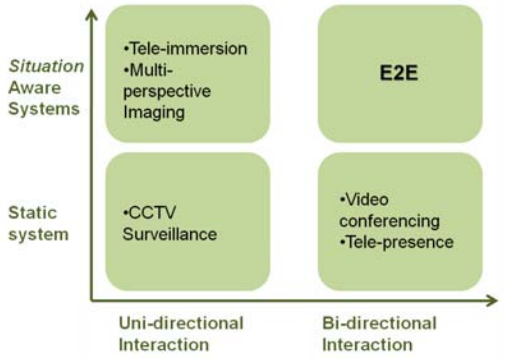


Figure 1: Comparison of emotion enhanced E2E with related works

are indeed the important triggers to re–compute the most appropriate sensors, actuators and their parameters.

Figure 2 shows a high–level architecture diagram of our EEE2E approach.

The 'Data acquisition' (DA) component gathers the relevant information from various sensors and undertakes the necessary processing on it. The data obtained is used to undertake emotion detection (ED) based on techniques described in section 2.1. The Environment Model (EM) creates linkages between the various sensors and their physical location. Thus if a camera and a microphone detect the sub–events of 'person present' and 'person angry', the environment model is useful in finding the location of this angry person. Any additional contextual information that may be required to convert these elemental level events[19] into application relevant events (e.g. student is angry with the professor) is added by the specific Situation Model (SM). The SM represents all domain–dependent information which is required to support application functionality.

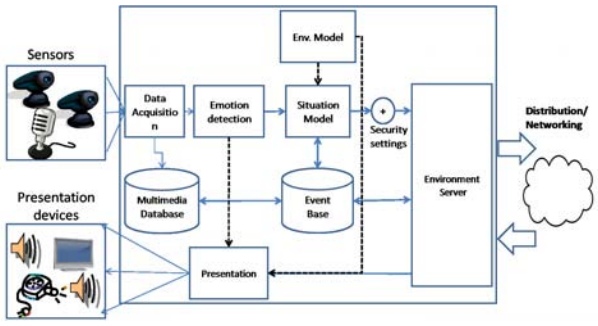


Figure 2: A high–level architecture diagram for EEE2E

The generated event are filtered based on the security and privacy settings before being put up on the Internet by the Event Server(ES). The ES will be responsible for routing out the most appropriate data streams (e.g. 'video stream(s)' and/or 'emotion status') as well as for routing the incoming data streams to be presented at most appropriate locations in conjunction with the presentation module. ES is also responsible for arbitrating and controlling incoming 'control requests' for available environment resources as well as for making such requests to other environments. All the gen-

erated multimodal information is archived in a multimedia database (MMDB), while the semantic level labels for all the events generated are stored in an Event–Base. The Event-Base does not store any media by itself but maintains links to relevant data in the MMDB.

The events (e.g. change from 'happy' to 'angry' emotion) act as triggers to initiate communication sessions across environments as well as to activate selection of appropriate sensors and presentation devices across environments. The actual distribution of the data is undertaken via peer–to–peer links over Internet between the various ESs. Each sensor and rendering device is seen as a web–service and is used by other Event Servers. The sharing of Event Servers over the Internet allows the users to collaborate across environments in their natural settings via virtualized 'Joint Situation Models'. The JSMs allow users opportunities to interact, collaborate and create new media and events which exist in totality only in the Joint Space.

2.1 Emotion Recognition

In our approach, emotion recognition is performed by fusing information coming from both visual and audio modalities. We are targeting the identification of the six "universal" emotions listed by Ekman and Friesen [20] (i.e. anger, disgust, fear, happiness, sadness, and fear).

According to the study of Ekman and Friesen these six emotions are characterized by the fact of being displayed via the same facial expression regardless of sex, ethnicity, age, and culture. As several researchers did before us [15], we implicitly make the assumption that these findings are true for emotional prosodic expression too.

The idea of using more than one modality arises from two main observations: 1) when one, or the other, modality is not available (e.g. the subject is silent or hidden from the camera) the system could still return an emotional estimation thanks to the other and 2) when both modalities are available, the diversity and complementarity of the information, should couple with an improvement on the general performances of the system.

2.1.1 Facial Expression Features

We have developed [21, 22] a system performing real time, user independent, emotional facial expression recognition from still pictures and video sequences.

Our technique makes use of keypoints (FP). We use several state of the art techniques such as the Viola–Jones detector [23], a 2D anthropometric model of the face (figure 3), and the Tomasi Lucas–Kanade algorithm [24]. These techniques allow us to extract 12 keypoints in real–time (see figure 4(a)).

The whole process could be described as follows.
The video is analyzed until a face is found using the Viola–Jones face detector [23]. We employ three detectors, one for the face, a second for the eyes, and a third for the mouth, in order to estimate the face orientation. Once the face position and its orientation have been estimated, we proceed to apply a simple two dimensional anthropometric model of the human face (see figure 3) similarly to what it was done in [25].

Thanks to this model, we can define 12 region of interest as in figure 3 corresponding to the following regions of the face (see also figure 4(a)):

1. right mouth corner
2. left mouth corner
3. nose
4. right eye
5. left eye
6. forehead
7. mouth bottom / chin
8. external right eyebrow
9. internal right eyebrow
10. internal left eyebrow
11. external left eyebrow
12. upper lip / mouth top

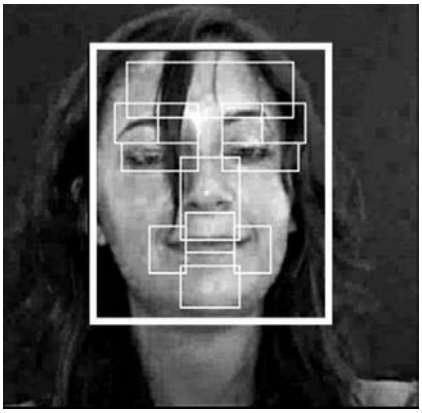


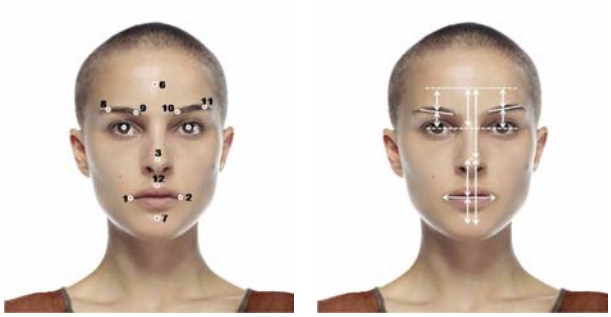
Figure 3: Anthropometric 2D model

For each one of these 12 regions we search for a cloud of points which will be easily tracked using the algorithm of Lucas–Kanade. We track these points all along the video and for each frame we compute the center of mass[2] as shown in figure 4(a). As a result of this process, we have 24 features per frame, corresponding to 12 pairs of the feature points (FP) $x(i)$ and $y(i)$ coordinates representing the average movement of points belonging to the regions of interest defined above.

We have also attempted to extract some more meaningful features, from these 24 coordinates (12 points times 2 dimensions), in a similar way to the one adopted by MPEG–4 Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs).This process resulted in 14 distances defined as follows (see also figure 4(b)):

- mouth corner distance
- chin distance to mouth
- nose distance to mouth
- nose distance to chin

[2]This last operation has been done to reduce a vibration effect that we did notice in our preliminary tests and which is mainly due to the compression blocking effect.



(a) Feature Points (b) Distances

Figure 4: Video Features

- left eye to eyebrow distance
- right eye to eyebrow distance
- left eyebrow alignment
- right eyebrow alignment
- left eyebrow to forehead distance
- right eyebrow to forehead distance
- forehead to eye line distance
- head x displacement
- head y displacement
- normalization factor proportional to head z displacement

Please note that all the processes needed to obtain this result have low computational requirements and can easily be used for real time processing. Furthermore, the process could easily be written for parallel computing and speed up 55X as described at www.nvidia.com/object/cuda/.

2.1.2 Prosodic Expression Features

Our system for speech emotion recognition, takes deep inspiration from the work of Noble [26]. From the audio signal we extract:

- the fundamental frequency or pitch (f_0)
- the energy of the signal (E)
- the first three formant (f_1, f_2, f_3)
- the harmonicity of the signal (HNR)
- the first ten linear predictive coding coefficients (LPC_1 to LPC_{10})
- the first ten mel–frequency cepstral coefficients ($MFCC_1$ to $MFCC_{10}$)

This sum up to 26 features which are collected with the use of PRAAT[3] [27] (www.praat.org/) and downsampled to 25 frame per second to help synchronization with video features. Also the processing time of the audio analysis is compatible with real–time constrains.

[3]PRAAT being a C++ toolkit written by P. Boersma and D. Weenink to record, process, and save audio signals and parameters.

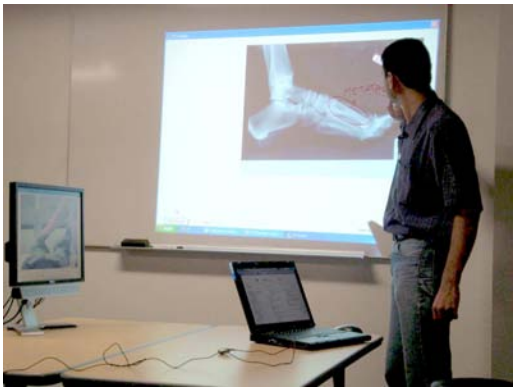


Figure 5: e–Learning scenario: the instructor environment

2.1.3 Classification and Post Processing

We used both support vector machines (SVM) and neural network (NN) to build different classifiers for audio and video. For each frame we have 6 likelihood corresponding to the probability that that particular frame, and the previous second of video[4], depict each one of the six universal emotion. Our tests on the eNTERFACE'05 database [28] show a correct recognition rate around of 45% which can be doubled to 90% by thresholding the results and taking only the best 4% of the evaluations (i.e. an average of one sample per second) [21, 22].

The eNTERFACE'05 database is composed of over 1300 emotionally tagged videos portraying non–native English speakers displaying a single emotion while verbalizing a semantically relevant English sentence. The 6 universal emotions from Ekman and Friesen [20] are portrayed. Videos have a duration ranging from 1.2 to 6.7 seconds (2.8 ± 0.8 sec). This database is publicly available on the Internet but carries few drawbacks due to the low quality of the video compression and actor performances (see [21] for a more extensive analysis of the database qualities).

3. EMOTION INTEGRATION IN E2E

[I have few doubts about this section and the next one]

Our current implementation of E2E relies on audio–visual content only. We used a total of 9 cameras, 5 microphones, 5 speakers and 6 display devices (1 multi–tiled display, 2 projectors, and 3 computer monitors) spread across the four different environments in two different continents. Three environments were set in Irvine (California) while the fourth was set in Sophia Antipolis (France). One PC in each environment acted as an Environment Server and undertook the necessary processing.

All the input and output devices were IP based (IP Cameras and IP microphones were realized using Axis Communication PTZ (Pan Tilt Zoom) duplex–audio support cameras). Epson 2315 IP–based projector and other Internet connected PC monitors were used to handle the display requirements. The use of IP based sensors/devices eased the implementation for the ES and also allowed the system to be scalable.

[4]When one second of video is not available, then the missing data is extrapolated from the accessible one.

We integrated the software for the recognition of the emotions on the Environment Server of each different location. For our first implementation of the emotion recognition module we focused on the sole video processing[5].

A simple software on the Event Servers is committed to detect which camera(s) better depict the faces in the environment and to pass this information to the emotion recognition module running on the environment server. Six likelihood are extracted for each subject and each video frame following the procedures briefly explained in section 2.1.1. We apply thresholding techniques to improve the quality of the emotion appraisal.

Emotions provide a semantically rich (yet achievable) abstraction of what is happening in each environment. Hence, the emotion feedback as shared by the Environment Server to the Event Server can be used to select the best sensor/actuators in both the environments. This decision will be made by a simple artificial intelligence and will depend upon the pair of specific emotion and scenario.

For example in a tele–education scenario, a "bored" expression from the student's environment conveys the need for instructor's environment to change the camera–feed angle, or use gaze–control and voice modulation etc. to engage him better. Furthermore, changing the camera–feed angle of the bored student's environment, may trigger the focus of the teacher who might decide upon changing the teaching stile.

4. EMOTIONS IN EDUCATION

Emotions play an intricate role in influencing how students learn in educational environment. In general it is well known that a certain level of attention, arousal, or even stress is needed to learn in the fastest way and perform better during exams [29]. Furthermore, previous research has pointed out that accurately identifying a learnerŠs emotional/cognitive state is a critical indicator of how to assist the learner in achieving an understanding of learning process [30].

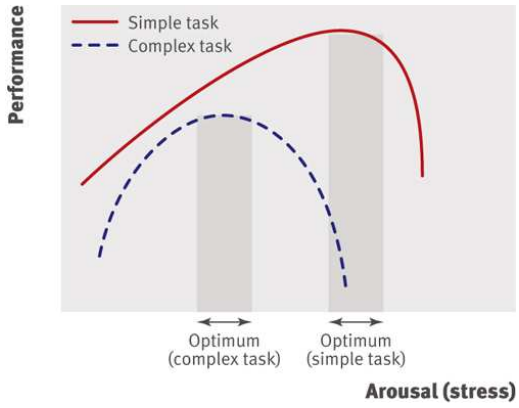


Figure 6: Yerkes–Dodson curves linking stress to learning performances

Expert educators are good at understanding such emotions and then reacting to the student learning needs. For example, if a student appears to be engaged in the task and enjoying trying things, even if he or she is making mistakes, then it might not be good to interrupt the student [31]. Thus in EEE2E framework we use student emotions to influence their learning activities.

There have been multiple attempts to study how different emotions are descriptive of the various learning phases which the student is undergoing. For example, Kort et al.[30] have described the different emotions which effect user learning and the related phases. A summary of these is shown in figure 7.

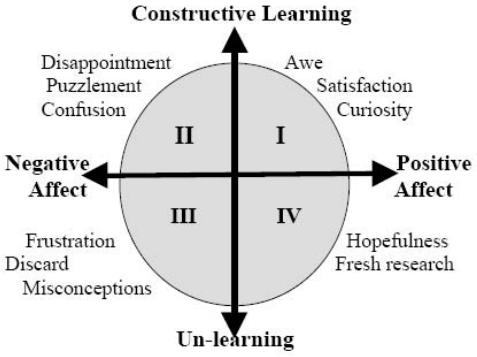


Figure 7: Impact of emotions on phases of user learning

5. EEE2E FOR EDUCATION

To ground our ideas on emotion-enhanced E2E and to verify its relevance to a remote education scenario we are implementing the following scenario. The instructor is a content expert with a classroom environment located in Irvine, California. The classroom consists of a projection based computer-screen for the instructor to teach, a white-board to derive and explain his mathematical ideas, a physical engineering model for some practical demonstrations and a multi-tiled display which displays remote users and their reactions to the instructor. The students are present in 3 other environments (2 in Irvine and 1 in Sophia-Antipolis, France) and have a basic Skype based audio-video (camera, microphone, speaker, monitor) connectivity with the remote classroom.

The instructor will spend most of his time standing near the projection with minor movements and lecturing about the subject (see figure 5). He intermittently could move to totally different locations (i.e. white-board and model) in the classroom. The most appropriate camera feed is selected and transmitted to the remote students as in our standard E2E development. However, based on the affective reactions of the various students the appropriate camera feed could be changed. This could be done for all the students, or could be done for some students depending on the percentage of students showing positive (and negative) emotional reactions to the lecture. The rule-base used for deciding on the appropriate camera to select was as follows.

```
$Std\_neg$= Percentage of students showing negative reactions
$Std\_pos$= Percentage of students showing positive reactions
$Time\_lastMove$= Time elapsed since last movement by
```

[5]In many scenarios multiple subject are interacting in the same environment. Then, a good diarization module it is needed to identify the subject speaking at a specific time.

	Influence Location	Influence Handler	Influence Type	Recent Event	Emotion Detected	Description
1	Remote student environment	System	Sensor selection	Lecturer movement	Small negative	If only few students want to relook the object at the previous location
2	Classroom environment	System	Sensor selection	Lecturer movement	Largely negative	If most students want to relook the object at the previous location
3	Classroom environment	System	Sensor positioning		Small negative	Lecturer not at center of the video capture. Small negative influence.
4	Classroom environment	User	Pedagogy method		Largely negative	If most students are "bored" with the lecture
5	Classroom environment	User	Pedagogy method		Small negative	If most students are "tired"

Table 1: Influence of user emotions on E2E communication

```
lecturer between locations.
$T_{thresh}$= 5 seconds
$Per_thres1$=0.25
$Per_thres2$=0.5

While (Time\_lastMove) < $T_{thresh}$
  IF ($Std\_neg$) > $Per_thres2$
   THEN SelectCamALL(PrevCam)
  IF ($Std\_neg$) < $Per_thres2$ and  ($Std\_neg$) > $Per_thres1$
   THEN SelectCamInd(PrevCam)
End While
  IF ($Std\_neg$) > $Per_thres2$
   THEN RepositionCam(CurrCam) and AlertInstructor(Alarm)
  IF ($Std\_neg$) > $Per_thres1$
   THEN RepositionCam(CurrCam)
```

As might be obvious from the above mentioned rule description, the use of emotions could influence the dynamics of camera tele-education in multiple ways. The different type of influence (on user, sensor selection and sensor positioning) has been summarized in table 1.

Furthermore two behavior are detected which shall influence the "pedagogy method". If students are tired or bored then appropriate student camera feed will help the teacher to acknowledge this situation and to change his/her teaching style.

6. OTHER SCENARIOS FOR EEE2E

6.1 Tele–medicine

Depression was 2004 third leading cause of disability in the world according to the 2008 World Health Organisation's "*Global Burden of Disease Report*" and previsions are that by year 2020 depression will be second only to heart disease.

Laughter and positive affective states help to combat stress and reduce pain by releasing the body's natural painkiller known as "*endorphins*" but it also have positive effects on the cardiovascular and respiratory systems, relax muscles, and boost the immune system by increasing the number of "*T–cells*" and lowering "*serum cortisol*" levels ([32]).

In tele–medicine, the doctor is interacting with patients at remote locations. While in face–to–face communications it is easy for the doctor to assess the emotional state of the patient and react accordingly, it might be hard to get the same information when communicating via a computer interface. With a EEE2E communications, not only the meeting will be more natural and then less stress–builder, but also the system could effectively and automatically inform the doctor of subtle modifications of the patient affective state. Assuming that our society likes to avoid disabilities, that monitoring of the affective state could help preventing depression to outcome, and that not everybody can afford private psychological follows–up, affective computing for the prevention of the depression state is going to become a more and more central topic of research in the next few years.

6.2 Instant Messaging and Tele–Conferencing

In our everyday lives we communicate more and more with remote locations thank to our computers and IM softwares. When communicating with remote environment the information about the emotions is usually lost. EEE2E have a natural application domain in this kind of scenario.

Indeed, while IM with our friends we often would like to also communicate a specific emotional state, or to clarify the meaning of a sentence with tag such as "ironic", "joke", or "sad". Although emoticons and smileys are quite useful improving the expressivity of online communication, they still provide a very limited means of expressing emotion. Furthermore, it remains within the responsibility of the user to carefully prepare the affective content of the textual message [33, 34]. EEE2E could help us in this kind of scenario by naturally and automatically include the correct emotional information into our messages.

6.3 Online Gaming

In the last decade the gaming industry has released several titles specifically designed for online gaming and opened new opportunities to compare our game skills with friends. Examples can be shoot–them–up games such as "Quake III Arena" and racing games such as "Virtual Skipper Online" but for sure the most interesting game domain is the one of massive multiplayer online games (MMOG) such as "Second Life" or "World of Warcraft". In all of these games the sense of interacting with other human being is central to the interest of the game itself. Specifically designed EEE2E systems could further improve this feeling by helping to naturally communicate in game with our friends.

7. CONCLUDING REMARKS

In this paper, we have described a new form of communication which supports natural human interaction by connecting environments to environments (E2E) rather than specific devices. We proposed an abstracted, event based, multimodal and scalable architecture to support such communications. We have shown that human everyday communications strongly rely on paralanguage. An important of this paralanguage is linked to emotions. Therefore, we described how emotions could be recognized with real–time technology and easily integrated in our E2E architecture to build up an emotion enhanced environment to environment (EEE2E) system. Then, we have given few simple examples on the possible use of emotions for tele–communications in the domain of tele–education and in other scenarios.

While we have described a successful initial implementation experience with EEE2E systems, there are multiple research challenges which need to be handled effectively for creation of sophisticated EEE2E systems. In particular we need to find reliable training data for the emotions "tired", "bored", "puzzled", "interested", and "challenged". Albeit we are able to partially overcome this lack of data by using a simple valence–arousal model, it is clear that this dedicated data would be beneficial to the precision and reliability of the emotion estimation.

Future work will focus on three main directions:

1. **emotion estimation refinement**.
   a) Some recent research of ours studied ways to better exploit the emotional information from single features (e.g. single distances among points, single coordinates, or single audio features)[]. We have hints that combining different processing for different features could improve the quality of the emotional estimates.
   b) We are working at making the software reliable for multimodal audio–visual emotion recognition. It is known [15] that exploiting multimodal data should improve both the results reliability and availability.

2. **behaviors development**.
   We have started to develop some simple reactions in the scenario of tele–education (e.g. camera–feed angle changes). More behaviors like this should be developed for different scenarios.

3. **user studies and quality assessment**.
   We are planning to assess the quality of the built system through some user studies.

8. REFERENCES

[1] Antonio R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Avon Books, NY, 1994.

[2] R. Picard, *Affective Computing*, MIT Press, Cambridge (MA), 1997.

[3] A. Mehrabian, *Nonverbal Communication*, Aldine-Atherton, 1972.

[4] D. Nguyen and J. Canny, "Multiview: Improving trust in group video conferencing through spatial faithfulness," in *Proceedings of CHI'07*, 2007, pp. 1465–1474.

[5] R. Vertegaal, G. Van der Veer, and H. Vons, "Effects of gaze on multiparty mediated communication," in *Proceedings of Graphics Interface*, 2000, pp. 95–102.

[6] B. Buxton A. Sellen and J. Arnott, "Using spatial cues to improve videoconferencing," in *Proceedings of CHI'92*, 1992, pp. 651–652.

[7] G. Mark and P. DeFlorio, "An experiment using life-size hdtv," in *Proceedings of IEEE Workshop on Advanced Collaborative Environments (WACE)*, 2001.

[8] G. Mark, S. Abrams, and N. Nassif, "Group-to-group distance collaboration: Examining the 'space between'," in *Proceedings of European Conference of Computer-supported Cooperative Work*, 2003, pp. 14–18.

[9] S. Bly, S. Harrison, and S. Irwin, "Media spaces: bringing people together in a video, audio, and computing environment," *Communications of the ACM*, vol. 36, no. 1, pp. 28–46, 1993.

[10] W. Gaver, T. Moran, A. MacLean, L. Lovstrand, P. Dourish, K. Carter, and W. Buxton, "Realizing a video environment: Europarc's rave system," in *Proceedings of CHI'92*, 1992, pp. 27–35.

[11] R. Stults, "Media space," Tech. Rep., Xerox PARC, 1986.

[12] Hewlett Packard, "Hp halo overview," 2007.

[13] SecondLife, "http://secondlife.com/," .

[14] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction," in *Proceedings of IEEE*, 2003, vol. 91, pp. 1370–1390.

[15] Z.Zeng, M. Pantic, G.I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, January 2009.

[16] C. L. Lisetti and C. LeRouge, "Affective computing in tele-home health: design science possibilities in recognition of adoption and diffusion issues," in *In Proceedings 37th IEEE Hawaii International Conference on System Sciences*, Hawaii, USA, 2004.

[17] Jeffrey W. Chastine, Kristine Nagel, Ying Zhu, and Luca Yearsovich, "Understanding the design space of referencing in collaborative augmented reality environments," in *GI '07: Proceedings of Graphics Interface 2007*, 2007, pp. 207–214.

[18] Patrick H. Kelly, Arun Katkere, Don Y. Kuramura, Saied Moezzi, and Shankar Chatterjee, "An architecture for multiple perspective interactive video," in *MULTIMEDIA '95: Proceedings of ACM International Conference on Multimedia*, 1995, pp. 201–212.

[19] Vivek K. Singh, Hamed Pirsiavash, Ish Rishabh, and Ramesh Jain, "Toward environment-to-environment (e2e) multimedia communication systems," in *SAME '08: Proceeding of the 1st ACM international workshop on Semantic ambient media experiences*, 2008, pp. 31–40.

[20] P. Ekman and W. V. Friesen, "A new pan cultural facial expression of emotion," *Motivation and Emotion*, vol. 10(2), pp. 159–168, 1986.

[21] M. Paleari and B. Huet, "Toward Emotion Indexing of Multimedia Excerpts," in *CBMI '08 Sixth International Workshop on Content-Based Multimedia Indexing*, London, June 2008, IEEE.

[22] M. Paleari, R. Benmokhtar, and B. Huet, "Evidence theory based multimodal emotion recognition," in *MMM '09* 15^th *International Conference on MultiMedia Modeling*, Sophia Antipolis, France, January 2009, ACM.

[23] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2001.

[24] C. Tomasi and T. Kanade, "Detection and tracking of point features," April 1991, CMU-CS-91-132.

[25] A.S.M. Sohail and P. Bhattacharya, *Signal Processing for Image Enhancement and Multimedia Processing*, vol. 31, chapter Detection of Facial Feature Points Using Anthropometric Face Model, pp. 189–200, Springer US, 2007.

[26] J. Noble, "Spoken emotion recognition with support vector machines," *PhD Thesis*, 2003.

[27] Paul Boersma and David Weenink, "Praat: doing phonetics by computer," January 2008, [http://www.praat.org/].

[28] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACEŠ05 Audio-Visual Emotion Database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006.

[29] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, pp. 459–482, 1908.

[30] B. Kort, R. Reilly, and R.W. Picard, "An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion," in *Proceedings of IEEE International Conference on Advanced Learning Technologies*, 2001, pp. 43–46.

[31] Ashish Kapoor and Rosalind W. Picard, "Multimodal affect recognition in learning environments," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 677–682.

[32] R. Martin, "Is laughter the best medicine? humor, laughter, and physical health," *Current Directions in Psychological Science*, vol. 11, pp. 217–219, 2002.

[33] M. Handel and J.D. Herbsleb, "What is chat doing in the workplace?," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2002, pp. 1–10, ACM Press.

[34] E. Isaacs, A. Walendowski, S. Whittaker, D.J. Schiano, and C. Kamm, "The character, functions, and styles of instant messaging in the workplace," in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, 2002, pp. 11–20, ACM Press.